26 BE 7082 + 26 PH 7028 + 20 BME 7082

Autumn 2020

MB Rao

Homework Sheet No. 7 Due Date: October 15, 2020 Maximum Points: 30

Wisconsin Breast Cancer Data: A gold standard procedure for detecting breast cancer is biopsy. This is a definitive procedure. Women, fifty years or older, are advised to be checked once every year for the presence or absence of cancer. Biopsy is painful, time-consuming, and expensive. It cannot be done every year. An alternative procedure is mammogram. This diagnostic test is not accurate. Its sensitivity is about 85% (True positives) and specificity 80% (True negatives). Wisconsin Medical Research Center proposed another diagnostic procedure (breast aspiration), which they touted as more accurate than the mammogram. A needle is inserted into the breast and cells are extracted. Various properties (nine in all) of the cells are noted for each case (malignant) and control (benign). Download the data (bopsy) from the package (MASS). The response variable is ‘class,’ which is binary. We have nine predictors in all.

Theme: Learn how to fit the logistic regression model to the data.

1. What is the dimension of the data? 1 point

**> dim(biopsy)**

**[1] 699 11**

2. Show the top ten rows of the data. 1 point

**> head(biopsy, 10)**

ID V1 V2 V3 V4 V5 V6 V7 V8 V9 class

1 1000025 5 1 1 1 2 1 3 1 1 benign

2 1002945 5 4 4 5 7 10 3 2 1 benign

3 1015425 3 1 1 1 2 2 3 1 1 benign

4 1016277 6 8 8 1 3 4 3 7 1 benign

5 1017023 4 1 1 3 2 1 3 1 1 benign

6 1017122 8 10 10 8 7 10 9 7 1 malignant

7 1018099 1 1 1 1 2 10 3 1 1 benign

8 1018561 2 1 2 1 2 1 3 1 1 benign

9 1033078 2 1 1 1 2 1 1 1 5 benign

10 1033078 4 2 1 1 2 1 2 1 1 benign

3. The first column is id. Create a new folder eliminating the first column. 1 point

**df<-subset(biopsy, select=-c(ID))**

4. What is the class of the first predictor? 1 point

**> class(df$V1)**

[1] "integer"

5. Obtain summary statistics of the data. Are there missing values? If there are, identify the variables in which they are located. 1 + 1 points

**> summary(df)**

V1 V2 V3

Min. : 1.000 Min. : 1.000 Min. : 1.000

1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000

Median : 4.000 Median : 1.000 Median : 1.000

Mean : 4.418 Mean : 3.134 Mean : 3.207

3rd Qu.: 6.000 3rd Qu.: 5.000 3rd Qu.: 5.000

Max. :10.000 Max. :10.000 Max. :10.000

V4 V5 V6

Min. : 1.000 Min. : 1.000 Min. : 1.000

1st Qu.: 1.000 1st Qu.: 2.000 1st Qu.: 1.000

Median : 1.000 Median : 2.000 Median : 1.000

Mean : 2.807 Mean : 3.216 Mean : 3.545

3rd Qu.: 4.000 3rd Qu.: 4.000 3rd Qu.: 6.000

Max. :10.000 Max. :10.000 Max. :10.000

NA's :16

V7 V8 V9

Min. : 1.000 Min. : 1.000 Min. : 1.000

1st Qu.: 2.000 1st Qu.: 1.000 1st Qu.: 1.000

Median : 3.000 Median : 1.000 Median : 1.000

Mean : 3.438 Mean : 2.867 Mean : 1.589

3rd Qu.: 5.000 3rd Qu.: 4.000 3rd Qu.: 1.000

Max. :10.000 Max. :10.000 Max. :10.000

class

benign :458

malignant:241

**There are 16 NA (missing) values.**

6. Describe the variables in the data. 4 points

**I used the following line of code for extracting information about the variables in data:**

**?biopsy**

This data frame contains the following columns:

ID

sample code number (not unique).

V1

clump thickness.

V2

uniformity of cell size.

V3

uniformity of cell shape.

V4

marginal adhesion.

V5

single epithelial cell size.

V6

bare nuclei (16 values are missing).

V7

bland chromatin.

V8

normal nucleoli.

V9

mitoses.

class

"benign" or "malignant".

7. Fit a logistic regression model to the data. Write the prediction model. Check goodness-of-fit. Identify the significant predictors. 3 + 3 + 3 + 3 points

I used the following line of code to fit the logistic regression model to the data:

**SF1<-glm(class~.,data=df, family=binomial)**

I used the following line of code to check the output:

**> summary(SF1)**

Call:

glm(formula = class ~ ., family = binomial, data = df)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.4841 -0.1153 -0.0619 0.0222 2.4698

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -10.10394 1.17488 -8.600 < 2e-16 \*\*\*

V1 0.53501 0.14202 3.767 0.000165 \*\*\*

V2 -0.00628 0.20908 -0.030 0.976039

V3 0.32271 0.23060 1.399 0.161688

V4 0.33064 0.12345 2.678 0.007400 \*\*

V5 0.09663 0.15659 0.617 0.537159

V6 0.38303 0.09384 4.082 4.47e-05 \*\*\*

V7 0.44719 0.17138 2.609 0.009073 \*\*

V8 0.21303 0.11287 1.887 0.059115 .

V9 0.53484 0.32877 1.627 0.103788

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 884.35 on 682 degrees of freedom

Residual deviance: 102.89 on 673 degrees of freedom

(16 observations deleted due to missingness)

AIC: 122.89

Number of Fisher Scoring iterations: 8

**The prediction equation is:**

I used the following lines of code for checking the goodness of fit:

**> pchisq(102.89, 678, lower.tail=F)**

**[1] 1**

**The fit is excellent as p-value 1>>0.05.**

Significant predictors are **V1, V6, V4, V7**. Insignificant predictors are V8, V2, V3, V5, V9.

8. Apply a model selection algorithm to get a tighter model. Write the prediction equation now. 2+1 points

I used the following lines of code for a tighter model:

**SF2<-glm(class~ V1+V4+V6+V7, data=df, family=binomial)**

**summary(SF2)**

Call:

glm(formula = class ~ V1 + V4 + V6 + V7, family = binomial, data = df)

Deviance Residuals:

Min 1Q Median 3Q Max

-3.6964 -0.1451 -0.0609 0.0232 2.4476

Coefficients:

Estimate Std. Error z value Pr(>|z|)

(Intercept) -10.11370 1.03264 -9.794 < 2e-16 \*\*\*

V1 0.81166 0.12585 6.450 1.12e-10 \*\*\*

V4 0.43412 0.11403 3.807 0.000141 \*\*\*

V6 0.48136 0.08816 5.460 4.76e-08 \*\*\*

V7 0.70154 0.15196 4.616 3.90e-06 \*\*\*

---

Signif. codes:

0 ‘\*\*\*’ 0.001 ‘\*\*’ 0.01 ‘\*’ 0.05 ‘.’ 0.1 ‘ ’ 1

(Dispersion parameter for binomial family taken to be 1)

Null deviance: 884.35 on 682 degrees of freedom

Residual deviance: 125.77 on 678 degrees of freedom

(16 observations deleted due to missingness)

AIC: 135.77

Number of Fisher Scoring iterations: 8

**pchisq(125.77, 678, lower.tail = F)**

**[1] 1**

**The fit is excellent.**

**The new prediction model:**

9. What happens to the missing observations when you fit the model to the data?

1 point

> SF2$na.action

24 41 140 146 159 165 236 250 276 293 295 298 316

24 41 140 146 159 165 236 250 276 293 295 298 316

322 412 618

322 412 618

attr(,"class")

**[1] "omit"**

**The missing observations are being omitted.**

10. Obtain the confusion matrix. 4 points

>install.packages(“regclass”)

>require(“regclass”)

**#confusion matrix for model containing all predictors.**

**> confusion\_matrix(SF1)**

Predicted benign Predicted malignant

Actual benign 434 10

Actual malignant 11 228

Total 445 238

**#confusion matric for model containing only significant predictors.**

**> confusion\_matrix(SF2)**

Predicted benign Predicted malignant

Actual benign 433 11

Actual malignant 13 226

Total 446 237

11. Calculate the misclassification rate. 1 points

**Misclassification rate of SF1** (original model with all predictors) = (11+10)/(445+238)=(21/683)=**0.030 or 3%**

**Misclassification rate of SF2** (containing only significant predictors) = (13+11)/(446+237)=(24/683)=**0.0351 or 3.51%**